CICA: Coupling Confidence-Aware Pretraining with Confidence-Informed Attention for Robust Multimodal Sentiment Analysis

Supplementary Material

A. Core Algorithms of CICA

To enhance reproducibility and clarify the key mechanisms within the CICA framework, this section presents the detailed pseudocode of two core components: (i) the training loop of CAP, and (ii) the forward propagation of the CIF module.

Algorithm 1: CAP

Algorithm 1 outlines the pretraining stage. The objective is to train each unimodal encoder E_m to extract features $\mathbf{H}_m$ while simultaneously learning to estimate the reliability of its own outputs. The encoder produces both feature embeddings and confidence-related signals, which are optimized through three complementary objectives.

Algorithm 1 Training loop of CAP

Require:
E_m: unimodal encoder
$\mathcal{D}_{\text{pretrain}}$: pretraining dataloader
θ: trainable parameters of E_m
$\lambda_{\text{CA}}, \lambda_{\text{uncert}}$: loss coefficients
$\mathcal{L}_{\text{task\_pre}}$: task loss (MSE)
$\mathcal{L}_{\text{CA}}$: knowledge alignment loss
$\mathcal{L}_{\text{uncert}}$: uncertainty regression loss
for each epoch **do**
  for each (batch_X, batch_y) in $\mathcal{D}_{\text{pretrain}}$ **do**
    $H_m, \hat{y}_m, s_m, u_m \leftarrow E_m(\text{batch\_X}, \text{is\_pretrain=True})$
    $L_{\text{task}} \leftarrow \mathcal{L}_{\text{task\_pre}}(\hat{y}_m, \text{batch\_y})$
    $L_{\text{conf}} \leftarrow \mathcal{L}_{\text{CA}}(s_m)$
    $u_{\text{target}} \leftarrow \tanh(|\text{batch\_y} - \hat{y}_m.\text{detach}()|)$
    $L_{\text{uncert}} \leftarrow \mathcal{L}_{\text{uncert}}(u_m, u_{\text{target}})$
    $\mathcal{L}_{\text{CAP}} \leftarrow L_{\text{task}} + \lambda_{\text{CA}} L_{\text{conf}} + \lambda_{\text{uncert}} L_{\text{uncert}}$
    ZERO_GRAD(θ) BACKWARD($\mathcal{L}_{\text{CAP}}$) STEP(θ)
  end for
end for
return trained encoder E_m

Algorithm 2: CIF

Algorithm 2 summarizes the forward propagation logic of the CIF module, which is used during multimodal fusion. The module first extracts structure-aware representations through a confidence-informed mechanism, followed by reliability-based modulation driven by the CAP outputs. This process corresponds to the two-step decoupled operation described in Sec. 3.1 of the main paper.

Algorithm 2 Forward propagation of the CIF module

Require:
1: Q: query vector (from $Z^{(l-1)}$)
2: $\{K_m, V_m\}_{m\in T,V,A}$: key–value pairs from encoders E_m
3: $\{s_m, u_m\}_{m\in T,V,A}$: confidence and uncertainty from CAP
4: $g(\cdot)$: projection that maps (s_m, u_m) to modulation signal r_m
5: CIA_{Attn}: structure-aware attention operator
6: $Z_{\text{outputs}} \leftarrow [\,]$
7: **for** each modality $m \in \{T, V, A\}$ **do**
8: $N_m \leftarrow \texttt{F.normalize}(\texttt{Conv1D}(K_m))$
9: $\rho_m \leftarrow \sigma(\texttt{Linear}_\rho(K_m))$
10: $S_{\text{geo}} \leftarrow \tanh(\texttt{einsum}(Q, N_m))$
11: $S_{\text{dens}} \leftarrow \texttt{Mean}(\rho_m)$
12: $S_{\text{mod}} \leftarrow w_g S_{\text{geo}} + w_d S_{\text{dens}}$
13: $A_m \leftarrow \texttt{softmax}(\frac{QK_m^\top}{\sqrt{d_k}} + S_{\text{mod}})$
14: $z_{\text{struct},m} \leftarrow A_m V_m$
15: $r_m \leftarrow \text{ReLU}(1 + s_m - u_m)$
16: $z_{\text{CIF},m} \leftarrow z_{\text{struct},m} \times r_m$
17: $Z_{\text{outputs}}.\text{append}(z_{\text{CIF},m})$
18: **end for**
19: $Z^{(l)} \leftarrow \texttt{LayerNorm}(Q + \sum Z_{\text{outputs}})$
20: **return** $Z^{(l)}$

The module first extracts structure-aware representations. As introduced in Sec. 3.2, this involves an **Intrinsic Structure Modulator (S_{mod})** to model the key's $\mathbf{K}_m$ intrinsic quality.The computation (Algorithm 2, lines 10-14) uses a 1D convolution to capture local contextual dependencies ($\mathbf{N}_m = \text{F.normalize}(\text{Conv1D}(\mathbf{K}_m))$) and a linear layer for token-level saliency ($\rho_m = \sigma(\text{Linear}_\rho(\mathbf{K}_m))$). These terms are aggregated (via S_{geo} and S_{dens}) into S_{mod}, which is then added to the attention matrix.This structure-aware representation ($z_{\text{struct},m}$) is then modulated by the reliability-based signal r_m (derived from CAP) to produce the final output $z_{\text{CIF},m}$.

B. Dataset and Evaluation Protocol Details

This appendix provides additionally information on the datasets, extracted features, and evaluation protocols used in all experiments presented in the main paper.

B.1. Dataset Details and Statistics

We evaluate the proposed CICA framework on four widely used multimodal sentiment analysis benchmarks: CH-SIMS [21], CH-SIMSv2 [10], MOSI [23], and MOSEI [2]. These datasets include both English and Chinese samples. For all experiments, we strictly follow the official train/validation/test splits to ensure full comparability with previous studies.

The four benchmarks provide a diverse evaluation setup. MOSI and MOSEI are English datasets annotated with continuous sentiment scores ranging from $[-3, 3]$. Their Chinese counterparts, CH-SIMS and CH-SIMSv2, offer similar contexts, though CH-SIMSv2 is notably larger (4403 vs. 2281 samples). The annotation schemes also differ; CH-SIMS uses a three-class system, whereas CH-SIMSv2 employs a finer-grained five-level annotation. Furthermore, CH-SIMSv2 adopts stricter cross-modal alignment, making it a more demanding testbed for evaluating robustness to modality misalignment and noise.

To ensure fair comparison and reproducibility, we adhere to established practices from prior studies [7, 22, 24] by using publicly released, pre-extracted features. These are provided in `.pkl` format and contain word-aligned feature sequences for all modalities. Specifically, for textual features, we use 768-dimensional embeddings from `bert-base-uncased` for MOSI/MOSEI and `bert-base-chinese` for CH-SIMS/CH-SIMSv2. For the audio and visual modalities, we adopt the standard acoustic feature sequences and official frame-level visual features supplied with each dataset. All features follow the common word-level alignment protocol, synchronized with the text sequence length.

B.2. Evaluation Protocol Details

We evaluate model performance and robustness under both regression and classification paradigms to ensure a comprehensive assessment.

For the **regression setting**, we adopt two standard metrics. We report Mean Absolute Error (MAE ↓), which measures the average L1 distance to reflect prediction deviation, and Pearson Correlation (Corr ↑), which quantifies the linear relationship between predicted and true sentiment values to indicate trend alignment.

For the **classification setting**, continuous sentiment scores are discretized following widely accepted practices. For binary classification (Acc-2, F1), two schemes are employed for MOSI and MOSEI: a *non-zero (non-0)* scheme where samples with $y > 0$ are Positive and $y < 0$ are Negative (excluding $y = 0$), and a *has-zero (has-0)* scheme where $y \geq 0$ is Non-negative and $y < 0$ as Negative. For CH-SIMS and CH-SIMSv2, a single division is used ($y \geq 0$ as Positive, $y < 0$ as Negative). For multiclass evaluation, labels are converted to discrete categories based on dataset conventions. For MOSI/MOSEI, we report 7-class accuracy (Acc-7) by rounding values in $[-3, 3]$ to the nearest integer. For CH-SIMS, we use its native three-class annotations (Negative, Neutral, Positive) and report Acc-3.

C. Implementation Details and Hyperparameters

Hardware and Software Our implementation is based on **PyTorch**. All experiments were conducted on a 32G cloud GPU (*e.g.* NVIDIA Tesla V100) using the AdamW optimizer.

Feature and Modality Alignment We follow standard practice [3, 14] and use publicly released, pre-extracted features for all modalities. For the textual modality, we use 768-dimensional embeddings from `bert-base-uncased` for MOSI/MOSEI and `bert-base-chinese` for CH-SIMS/CH-SIMSv2. Following [14], we align the visual and acoustic modalities to the text sequence length using a 1D convolutional network.

Training Procedure and Hyperparameters As described in the main text, training follows a two-phase procedure.

- **Phase 1 (CAP):** We pretrain the unimodal encoders (Sec. 3.1) for **50 epochs** using the full $\mathcal{L}_{\text{CAP}}$ objective. The learning rate for this phase is set to $\mathbf{1.0 \times 10^{-4}}$.
- **Phase 2 (CIF):** We freeze the pretrained encoders and fine-tune the CIF fusion module (Sec. 3.2) and the final prediction head for **100 epochs**. The learning rate for this finetuning phase is set to $\mathbf{5.0 \times 10^{-5}}$.

The MCP loss weight λ_{mcp} (Sec. 3.3) was set to **0.1**, determined via grid search on the validation set.

Tab. 5 lists the key hyperparameter settings used in our experiments.

D. Hyperparameter Sensitivity Analysis

We examine the sensitivity of CICA to its three main hyperparameters: the loss weights for Confidence Alignment (λ_{CA}) and Uncertainty Prediction (λ_{uncert}) in the CAP phase, and the loss weight for Mutual-information Contrastive Preservation (λ_{mcp}) in the fusion phase. Each parameter is varied independently while keeping the others fixed at their optimal settings, with results summarized in Tab. 8, Tab. 9, and Tab. 10.

The results reveal consistent trends across all four datasets. Within the CAP phase, disabling either $\mathcal{L}_{\text{CA}}$ or $\mathcal{L}_{\text{uncert}}$ leads to the largest performance degradation. For example, on CH-SIMS, removing $\mathcal{L}_{\text{CA}}$ reduces the correlation from 0.754 to 0.621, while omitting $\mathcal{L}_{\text{uncert}}$ causes a similar drop to 0.625. This finding directly supports

Table 5. Key Hyperparameter Settings.

Phase	Hyperparameter	Value
General	Optimizer	AdamW
General	Batch Size	32
General	Weight Decay	0.01
Phase 1 (CAP)	Pretrain Epochs	50
Phase 1 (CAP)	Learning Rate	1.0×10^{-4}
Phase 1 (CAP)	Encoder Layers (N_E)	4
Phase 1 (CAP)	Encoder Heads (H_E)	8
Phase 1 (CAP)	λ_{CA} (Confidence Loss)	0.5
Phase 1 (CAP)	λ_{uncert} (Uncertainty Loss)	0.5
Phase 2 (CIF)	Finetune Epochs	100
Phase 2 (CIF)	Learning Rate	5.0×10^{-5}
Phase 2 (CIF)	λ_{mcp} (MCP Loss)	0.1
Phase 2 (CIF)	Fusion Model Dim (D_{model})	512
Phase 2 (CIF)	Fusion Layers (N_F)	3
Phase 2 (CIF)	CIF Heads (H_F)	8
Phase 2 (CIF)	CIF Kernel Size (k_{norm})	5

our perceive–decide hypothesis: the reliability signals (s_m and u_m) learned through these objectives are essential for the confidence-informed coupling mechanism (Eq. (11)). When either term is removed, the encoders $\mathbf{E}_m$ fail to form meaningful reliability estimates, and the resulting noise propagation severely disrupts decision-level fusion. The model achieves its best performance when both loss weights are set to moderate values ($\lambda_{CA} = \lambda_{uncert} = 0.5$); larger values (*e.g.* 1.0 or 2.0) slightly reduce accuracy, suggesting that overemphasizing these auxiliary tasks begins to interfere with representation learning.

A similar pattern is observed for the MCP regularization term in the fusion phase. Disabling $\mathcal{L}_{\text{MCP}}$ ($\lambda_{mcp} = 0$) leads to a clear decline across all benchmarks—for instance, on MOSI, the F1 (non-0) score drops from 90.23 to 87.66—indicating that this constraint effectively prevents modality collapse. Optimal performance is obtained with a small weight ($\lambda_{mcp} = 0.1$), which maintains information preservation without over-regularizing the fusion process. Increasing the value further slightly harms performance, implying that $\mathcal{L}_{\text{MCP}}$ serves best as a stabilizing term rather than a primary optimization objective.

Overall, the framework exhibits stable performance across a wide range of hyperparameter values (approximately 0.1–1.0 for CAP losses), and the chosen configuration ($\lambda_{CA} = 0.5$, $\lambda_{uncert} = 0.5$, $\lambda_{mcp} = 0.1$) achieves the best balance between stability and discriminability. Importantly, setting any of these weights to zero causes substantial degradation, confirming that all three components are indispensable to the effectiveness of CICA.

E. Full Ablation study

This section provide the complete ablation study, which includes the additional results on the Chinese datasets, CH-SIMS and CH-SIMSv2. The comprehensive results across all four benchmarks are detailed in Tab. 6.

As shown in Tab. 6, the performance trends observed on the English datasets strictly hold for the Chinese datasets, demonstrating the strong cross-lingual generalizability of the proposed modules. Specifically, the removal of CAP (A) or CIF (B) leads to the most severe performance degradation on CH-SIMS (*e.g.*, Corr drops by 0.144 and 0.124, respectively) and CH-SIMSv2 (*e.g.*, Acc-3 drops by 4.53 and 3.48, respectively). This echoes our findings in the main text that confidence-aware pretraining and cross-modal interaction are fundamental to the framework's success.

Furthermore, disabling the explicit CAP-CIF coupling (D) causes a notable decline (*e.g.*, CH-SIMS Acc-3 drops by 4.34), confirming that these modules must operate synergistically rather than independently. Finally, removing MCP (C) or the intrinsic term S_{mod} (E) results in minor but consistent drops across all metrics on the Chinese datasets, reaffirming their roles in maintaining modality balance and preventing representation collapse regardless of the language domain.

F. Extended Robustness Analysis

This section provides the complete empirical results supporting the robustness study in the main paper (Sec. 4.5). Tab. 11 reports the detailed outcomes of modality noise injection across all four benchmarks, complementing the visualization in Fig. 3. Tab. 12 extends the missing-modality evaluation (Tab. 4 in the main text) to include both CH-SIMS and CH-SIMSv2.

F.1. Comparative Analysis on Missing-Modality Robustness

To complement the robustness results presented in Sec. 4.5, we further compare CICA with several recent baselines (KuDA[3], ALMT[24], CubeMLP[13]) under missing-modality conditions. The experiment, summarized in Tab. 7, evaluates model stability on the MOSI dataset when one or more input modalities are deliberately removed at inference.

Overall, the results highlight three consistent patterns. First, the (V+A) configuration again confirms the central role of text in sentiment reasoning. Removing textual input leads to a substantial degradation across all methods, echoing prior observations in Sec. 4.5.

Second, under partial conditions such as (T+V) or (T+A), CICA shows notably higher resilience. When the audio stream is absent (T+V), CICA reaches a correlation of **0.839**, outperforming KuDA (0.770), ALMT (0.778), and

Table 6. Ablation study of the CICA framework on four datasets. Columns represent ablated variants: **(Full)** complete model; **(A)** w/o CAP; **(B)** w/o CIF; **(C)** w/o MCP; **(D)** w/o Coupling; **(E)** w/o S_{mod}. For MOSI/MOSEI, results follow standard Acc-2/F1 under both has-0 and non-0 schemes. Δ denotes the absolute performance drop. Best results are in **bold**.

Metric	CICA (Full)	(A) w/o CAP	(B) w/o CIF	(C) w/o MCP	(D) w/o Coupling	(E) w/o S_{mod}
Dataset: MOSI						
MAE ↓	**0.630**	0.712 (↑0.089)	0.689 (↑0.066)	0.635 (↑0.012)	0.658 (↑0.035)	0.641 (↑0.018)
Corr ↑	**0.855**	0.791 (↓0.064)	0.812 (↓0.043)	0.847 (↓0.008)	0.831 (↓0.024)	0.840 (↓0.015)
Acc-7 ↑	**49.56**	46.47 (↓3.09)	47.52 (↓2.04)	49.13 (↓0.43)	48.25 (↓1.31)	48.82 (↓0.74)
Acc-2 (has-0) ↑	**88.19**	85.12 (↓3.07)	86.03 (↓2.16)	87.84 (↓0.35)	87.08 (↓1.11)	87.55 (↓0.64)
Acc-2 (non-0) ↑	**90.24**	86.88 (↓3.36)	87.90 (↓2.34)	89.51 (↓0.73)	88.73 (↓1.51)	89.30 (↓0.94)
F1 (has-0) ↑	**88.14**	84.03 (↓4.11)	85.21 (↓2.93)	87.05 (↓1.09)	86.53 (↓1.61)	87.48 (↓0.66)
F1 (non-0) ↑	**90.23**	85.05 (↓5.18)	86.13 (↓4.10)	87.66 (↓2.57)	87.02 (↓3.21)	89.15 (↓1.08)
Dataset: MOSEI						
MAE ↓	**0.489**	0.524 (↑0.035)	0.511 (↑0.022)	0.496 (↑0.007)	0.504 (↑0.015)	0.499 (↑0.010)
Corr ↑	**0.856**	0.775 (↓0.081)	0.793 (↓0.063)	0.845 (↓0.011)	0.820 (↓0.036)	0.836 (↓0.020)
Acc-7 ↑	**55.29**	52.81 (↓2.48)	53.54 (↓1.75)	54.82 (↓0.47)	54.17 (↓1.12)	54.50 (↓0.79)
Acc-2 (has-0) ↑	**84.72**	82.04 (↓2.68)	82.87 (↓1.85)	84.06 (↓0.66)	83.51 (↓1.21)	83.95 (↓0.77)
Acc-2 (non-0) ↑	**90.18**	87.51 (↓2.67)	88.04 (↓2.14)	89.42 (↓0.76)	88.91 (↓1.27)	89.43 (↓0.75)
F1 (has-0) ↑	**85.15**	82.52 (↓2.63)	83.10 (↓2.05)	84.11 (↓1.04)	83.82 (↓1.33)	84.35 (↓0.80)
F1 (non-0) ↑	**90.16**	83.19 (↓6.97)	83.90 (↓6.26)	84.71 (↓5.45)	84.13 (↓6.03)	88.05 (↓2.11)
Dataset: CH-SIMS						
MAE ↓	**0.378**	0.415 (↑0.037)	0.405 (↑0.027)	0.385 (↑0.007)	0.392 (↑0.014)	0.387 (↑0.009)
Corr ↑	**0.754**	0.610 (↓0.144)	0.630 (↓0.124)	0.735 (↓0.019)	0.690 (↓0.064)	0.724 (↓0.030)
Acc-3 ↑	**76.37**	66.08 (↓10.29)	68.12 (↓8.25)	74.06 (↓2.31)	72.03 (↓4.34)	74.11 (↓2.26)
Acc-2 ↑	**86.00**	81.84 (↓4.16)	82.53 (↓3.47)	85.31 (↓0.69)	84.54 (↓1.46)	85.20 (↓0.80)
F1 ↑	**85.68**	81.52 (↓4.16)	82.04 (↓3.64)	85.13 (↓0.55)	84.07 (↓1.61)	84.95 (↓0.73)
Dataset: CH-SIMSv2						
MAE ↓	**0.245**	0.275 (↑0.030)	0.268 (↑0.023)	0.250 (↑0.005)	0.258 (↑0.013)	0.253 (↑0.008)
Corr ↑	**0.842**	0.750 (↓0.092)	0.770 (↓0.072)	0.830 (↓0.012)	0.805 (↓0.037)	0.820 (↓0.022)
Acc-3 ↑	**80.56**	76.03 (↓4.53)	77.08 (↓3.48)	79.52 (↓1.04)	78.54 (↓2.02)	79.35 (↓1.21)
Acc-2 ↑	**85.98**	82.31 (↓3.67)	83.12 (↓2.86)	85.44 (↓0.54)	84.72 (↓1.26)	85.05 (↓0.93)
F1 ↑	**85.89**	81.84 (↓4.05)	82.51 (↓3.38)	85.21 (↓0.68)	84.33 (↓1.56)	85.10 (↓0.79)

CubeMLP (0.745) under the same setting.

Finally, even when operating with only two modalities, CICA maintains an advantage over all baselines that rely on the complete three-modality input. For instance, CICA's (T+V) and (T+A) settings achieve 0.839 and 0.840 correlation respectively—both exceeding the full-modality scores of KuDA (0.795), ALMT (0.793), and CubeMLP (0.772).

These results provide direct empirical support for CICA's adaptive "perceive–decide" mechanism. The model not only achieves competitive accuracy under ideal conditions but also retains substantially higher stability and absolute performance when information from one or more modalities is missing.

F.2. Resilience to Modality Noise

As summarized in Tab. 11, Gaussian noise ($\sigma \in [0.2, 0.8]$) was applied to the audio (A), visual (V), or combined (A+V) streams. The results reveal two clear patterns. First, visual noise consistently produces the most severe degradation. At $\sigma = 0.8$, correlation nearly collapses on all datasets (*e.g.* from 0.855 to 0.013 on MOSI, and from 0.842 to -0.406 on CH-SIMSv2). This indicates that visual cues carry the highest predictive value within the multimodal context; once corrupted, the model effectively loses access to one of its primary semantic anchors.

In contrast, the effect of audio noise is more dataset-dependent. On CH-SIMS, even strong audio corruption has negligible impact, while on MOSI it leads to a dramatic performance drop (Corr: $0.855 \rightarrow 0.100$). This divergence aligns with the behavior predicted by our perceive–decide framework. During pretraining, the CAP module learns modality reliability: for CH-SIMS, audio is intrinsically noisy and is assigned low confidence (s_A), so injecting further noise barely affects the final decision. For MOSI, audio is highly informative and thus strongly weighted; when perturbed, the model detects its degradation (confidence s_A decreases) and downweights it, but the resulting loss of this critical cue inevitably harms performance. These observa-

Table 7. Comparative robustness analysis on MOSI under missing modality conditions. We compare CICA against SOTA baselines.

Methods	MOSI	
	MAE	Corr
CICA	**0.630**	**0.855**
V+A	1.442	0.247
T+V	0.663	0.839
T+A	0.695	0.840
KuDA	0.705	0.795
V+A	1.37	0.235
T+V	0.769	0.77
T+A	0.733	0.795
ALMT	0.712	0.793
V+A	1.437	0.201
T+V	0.772	0.778
T+A	0.736	0.788
CubeMLP	0.755	0.772
V+A	1.453	0.137
T+V	0.796	0.745
T+A	0.816	0.739

tions confirm that CICA adapts appropriately to perceived reliability, rather than failing to respond to noise.

Tab. 12 further evaluates how CICA behaves when one or more input modalities are unavailable. The absence of the textual stream leads to the sharpest degradation across all datasets (*e.g.* Corr: $0.754 \rightarrow 0.014$ on CH-SIMS), which reaffirms that language remains the primary carrier of sentiment information. In contrast, when non-textual inputs are missing, the model remains notably stable. On MOSEI, removing the visual modality yields a Corr of 0.850, nearly matching the full configuration (0.856). This robustness can be attributed to the CAP module's ability to recognize missing inputs ($s_V = 0$) and to the CIA module's subsequent rebalancing of attention toward the remaining reliable signals.

A result from our robustness evaluation (Tab. 12) warrants attention: on the MOSI dataset, the text-only (T-only) configuration attains a correlation of 0.857, which slightly surpasses the full T+V+A model (0.855). This minor discrepancy does not suggest a weakness of the CICA framework. Rather, it reflects an intrinsic property of MOSI and reinforces the rationale of our perceive-and-decide design. The observation can be explained by three dataset-specific factors. First, as shown in our unimodal results and prior studies [7, 24], MOSI is unusually text-centric: its visual and acoustic channels are often noisy or contradictory to sentiment labels. Consequently, textual cues dominate prediction performance. The T-only model thus benefits from using this single, highly reliable modality.

In the full T+V+A configuration, CICA's CAP module (Sec. 3.1) correctly assigns near-zero reliability to the V and A inputs (i.e., $s_m \approx 0$, $u_m \approx 1$). Our coupling function (Eq. (10)) correctly interprets this and assigns a suppression weight of zero: $r_m = \text{ReLU}(1 + 0 - 1) = 0$. This suppression, however, is countered by the $\mathcal{L}_{\text{MCP}}$ objective (Sec. 3.3), which is part of the fine-tuning loss $\mathcal{L}_{\text{Total}}$. $\mathcal{L}_{\text{MCP}}$ is designed to prevent complete modality collapse by encouraging the fused representation $\mathbf{z}_{\text{final}}$ to retain mutual information with all unimodal sources. During fine-tuning, the model thus balances two objectives: the coupling function (Eq. 10) attempts to zero out the noisy V/A modalities, while the $\mathcal{L}_{\text{MCP}}$ loss acts as a key regularizer to ensure the fused representation retains sufficient multimodal information, thus preventing modality collapse and preserving generalizability. The final model parameters are a compromise. To satisfy the $\mathcal{L}_{\text{MCP}}$ constraint, the model must integrate a small amount of information from the V/A streams. On MOSI, where these cues are "noisy or contradictory", this mandated inclusion of noisy signals acts as the "slight 'fusion cost'," marginally perturbs the dominant text signal and yields the 0.002 performance difference.

G. Full Visualization of Learned Representations

Following the t-SNE analysis in Sec. 4.7, we present here the complete visualization results and corresponding quantitative cluster statistics for all four benchmarks. Each plot projects the final fusion-layer embeddings from the CIF module into a two-dimensional space, enabling examination of the topological structure of the learned representations.

Fig. 6 shows that across all datasets, CICA produces a smooth and semantically organized sentiment manifold. The representations form a continuous gradient from negative (blue) to positive (red), with neutral samples (green/gray) naturally positioned between the two extremes. This progression indicates that the fusion module captures not only categorical distinctions but also the subtle continuum of sentiment.

To quantify these observations, we compute inter-cluster separation and intra-cluster cohesion from the t-SNE embeddings. The learned space exhibits strong discriminability: the centroid distance between Positive and Negative samples consistently forms the widest gap, far exceeding their respective distances to the Neutral cluster (*e.g.* CH-SIMSv2: 66.20 vs. 31.86; MOSEI: 58.46 vs. 42.81; CH-SIMS: 40.74 vs. 21.27; MOSI: 35.76 vs. 23.66). At the same time, cluster cohesion remains tight, particularly for the polar categories—on CH-SIMS, for instance, the average spread for Positive and Negative clusters is only 11.43 and 12.13, while the Neutral class shows higher variability (23.86 on CH-SIMSv2), which aligns with its inherently diverse semantics.

Table 8. Sensitivity analysis for the loss weight λ_{CA}. The **Acc-7/3** column refers to Acc-7 for MOSI/MOSEI and Acc-3 for CH-SIMS/CH-SIMSv2. ↓ indicates lower is better, ↑ indicates higher is better. Optimal results are in **bold**.

Param.	Value	Dataset	MAE ↓	Corr ↑	Acc-7/3 ↑	Acc-2 ↑	F1 ↑
λ_{CA}	0	MOSI	0.671	0.823	47.81	86.95 / 88.10	86.21 / 87.90
	0.1	MOSI	0.640	0.845	48.92	87.73 / 89.44	87.01 / 89.12
	0.5	MOSI	**0.630**	**0.855**	**49.56**	**88.19 / 90.24**	**88.14 / 90.23**
	1.0	MOSI	0.630	0.851	49.21	88.01 / 89.95	87.80 / 89.73
	2.0	MOSI	0.638	0.846	48.87	87.65 / 89.60	87.14 / 89.10
	0	CH-SIMS	0.410	0.621	67.12	82.01	81.77
	0.1	CH-SIMS	0.389	0.725	73.50	85.15	84.90
	0.5	CH-SIMS	**0.378**	**0.754**	**76.37**	**86.00**	**85.68**
	1.0	CH-SIMS	0.381	0.748	75.80	85.70	85.42
	2.0	CH-SIMS	0.384	0.741	75.11	85.45	85.03
	0	MOSEI	0.519	0.780	53.05	82.41 / 87.88	82.88 / 83.45
	0.1	MOSEI	0.498	0.840	54.55	83.90 / 89.30	84.30 / 89.20
	0.5	MOSEI	**0.489**	**0.856**	**55.29**	**84.72 / 90.18**	**85.15 / 90.16**
	1.0	MOSEI	0.493	0.851	55.01	84.41 / 89.90	84.81 / 89.85
	2.0	MOSEI	0.497	0.847	54.76	84.15 / 89.71	84.40 / 89.42
	0	CH-SIMSv2	0.271	0.759	76.44	82.70	82.15
	0.1	CH-SIMSv2	0.254	0.821	79.01	85.10	84.90
	0.5	CH-SIMSv2	**0.245**	**0.842**	**80.56**	**85.98**	**85.89**
	1.0	CH-SIMSv2	0.248	0.838	80.15	85.70	85.60
	2.0	CH-SIMSv2	0.251	0.835	79.80	85.51	85.35

Table 9. Sensitivity analysis for the loss weight λ_{uncert}. Optimal results are in **bold**.

Param.	Value	Dataset	MAE ↓	Corr ↑	Acc-7/3 ↑	Acc-2 ↑	F1 ↑
λ_{uncert}	0	MOSI	0.664	0.826	48.02	87.11 / 88.30	86.44 / 88.05
	0.1	MOSI	0.638	0.848	49.01	87.80 / 89.60	87.10 / 89.20
	0.5	MOSI	**0.630**	**0.855**	**49.56**	**88.19 / 90.24**	**88.14 / 90.23**
	1.0	MOSI	0.629	0.852	49.30	88.05 / 90.01	87.90 / 89.81
	0	CH-SIMS	0.408	0.625	67.53	82.21	81.90
	0.1	CH-SIMS	0.386	0.730	73.51	85.41	85.15
	0.5	CH-SIMS	**0.378**	**0.754**	**76.37**	**86.00**	**85.68**
	1.0	CH-SIMS	0.380	0.750	76.02	85.83	85.50
	0	MOSEI	0.515	0.788	53.20	82.60 / 88.00	83.00 / 83.60
	0.1	MOSEI	0.495	0.843	54.70	84.10 / 89.50	84.50 / 89.40
	0.5	MOSEI	**0.489**	**0.856**	**55.29**	**84.72 / 90.18**	**85.15 / 90.16**
	1.0	MOSEI	0.492	0.853	55.10	84.50 / 90.00	84.90 / 89.90
	0	CH-SIMSv2	0.269	0.765	76.84	82.91	82.30
	0.1	CH-SIMSv2	0.252	0.825	79.32	85.30	85.00
	0.5	CH-SIMSv2	**0.245**	**0.842**	**80.56**	**85.98**	**85.89**
	1.0	CH-SIMSv2	0.247	0.840	80.31	85.80	85.70

Overall, both qualitative visualization and quantitative measurements confirm that CICA learns a feature space that is simultaneously separable and topologically ordered, providing a coherent explanation for its robustness and superior fusion behavior.

H. Justification for Core Methodological Components

H.1. Justification for the Adaptive CA Loss

A key component of the CAP stage is the CA loss ($\mathcal{L}_{\text{CA}}$). The need for a learnable, adaptive objective, rather than simpler static formulations such as Binary Cross-Entropy

Table 10. Sensitivity analysis for the loss weight λ_{mcp}. Optimal results are in **bold**.

Param.	Value	Dataset	MAE ↓	Corr ↑	Acc-7/3 ↑	Acc-2 ↑	F1 ↑
λ_{mcp}	0	MOSI	0.635	0.847	49.13	87.84 / 89.51	87.05 / 87.66
	0.05	MOSI	0.628	0.851	49.30	88.05 / 89.90	87.90 / 89.70
	0.1	MOSI	**0.630**	**0.855**	**49.56**	**88.19 / 90.24**	**88.14 / 90.23**
	0.2	MOSI	0.627	0.852	49.35	88.00 / 89.95	87.85 / 89.75
	0.5	MOSI	0.633	0.848	49.05	87.75 / 89.40	87.15 / 89.00
	0	CH-SIMS	0.385	0.735	74.06	85.31	85.13
	0.05	CH-SIMS	0.380	0.748	75.21	85.73	85.40
	0.1	CH-SIMS	**0.378**	**0.754**	**76.37**	**86.00**	**85.68**
	0.2	CH-SIMS	0.379	0.751	76.15	85.85	85.55
	0.5	CH-SIMS	0.383	0.740	75.02	85.41	85.20
	0	MOSEI	0.496	0.845	54.82	84.06 / 89.42	84.11 / 84.71
	0.05	MOSEI	0.491	0.851	55.05	84.40 / 89.80	84.80 / 89.70
	0.1	MOSEI	**0.489**	**0.856**	**55.29**	**84.72 / 90.18**	**85.15 / 90.16**
	0.2	MOSEI	0.492	0.853	55.15	84.50 / 90.00	84.95 / 89.90
	0.5	MOSEI	0.497	0.847	54.90	84.10 / 89.50	84.20 / 89.00
	0	CH-SIMSv2	0.250	0.830	79.52	85.44	85.21
	0.05	CH-SIMSv2	0.247	0.838	80.12	85.70	85.50
	0.1	CH-SIMSv2	**0.245**	**0.842**	**80.56**	**85.98**	**85.89**
	0.2	CH-SIMSv2	0.246	0.840	80.41	85.80	85.70
	0.5	CH-SIMSv2	0.249	0.833	79.90	85.50	85.30

(BCE) or Focal Loss, stems from the nature of reliability modeling. This task is fundamentally a **regression and calibration problem**, not a binary classification one.

Static classification losses such as BCE or Focal Loss are unsuitable as they frame reliability estimation as a binary task ($s_m \rightarrow \{0, 1\}$). This necessitates creating pseudo-labels from an arbitrary error threshold (*e.g.* "error ¡ 0.5 is reliable"). Such an approach is ill-posed; it introduces dataset-specific hyperparameters that generalize poorly and enforces a rigid decision boundary where reliability actually varies continuously.

While a static regression objective, such as $\mathcal{L} = (s_m - 1)^2$, avoids this binary framing, it remains deficient. It statically assumes a uniform penalty for underconfidence (*e.g.* $s_m = 0.7$) across all samples and datasets. However, the semantics of confidence are highly context- and dataset-dependent.

$\mathcal{L}_{\text{CA}}$ resolves these issues by functioning as a **learnable calibration function**. It acknowledges that the meaning of "high confidence" varies; a score of $s_m = 0.7$ might indicate certainty in a clean corpus but low reliability amid heavy modality conflict. By introducing learnable parameters $\theta_{\text{CA}} = \{\hat{\alpha}, \hat{\beta}, \hat{w}_{\text{high}}, \dots\}$, $\mathcal{L}_{\text{CA}}$ allows the model to infer dataset-specific reliability boundaries directly from the data. Instead of applying a pre-defined curve (such as BCE or static L2), it learns adaptive boundaries for "high," "mid," and "low" reliability, which is evidenced by stable, distinct convergence patterns across different datasets.

Thus, $\mathcal{L}_{\text{CA}}$ transforms the encoder from a conventional feature extractor into an **adaptively calibrated perceiver**, which enables robust performance under heterogeneous noise and modality conflict.

H.2. Empirical Analysis of Signal Complementarity

A core premise of our perceive-and-decide framework is that the adaptive confidence (s_m) and predicted uncertainty (u_m) signals, both generated by the CAP module, are complementary, not redundant. We provide empirical validation for this claim by re-interpreting the hyperparameter sensitivity analysis from Appendix D as a ablation study. Setting the loss weights λ_{CA} or λ_{uncert} to zero effectively removes the corresponding signal component from the model.

Setting $\lambda_{uncert} = 0$ forces the model to rely solely on the adaptive confidence s_m. As shown in Tab. 9, this results in a severe performance drop across all datasets (*e.g.* MOSI Corr: 0.855 → 0.826; CH-SIMS Corr: 0.754 → 0.625). This demonstrates that s_m alone is insufficient. A modality can be "confident but wrong"—such as a confident smile during a sarcastic utterance—and the u_m signal is required to penalize this high, unperceived objective error. Conversely, setting $\lambda_{CA} = 0$ removes the adaptive confidence signal, relying only on the predicted error u_m. This ablation also triggers a catastrophic performance collapse (*e.g.* MOSI Corr: 0.855 → 0.823; CH-SIMS Corr: 0.754 → 0.621), proving that u_m alone is likewise insufficient.

The severe performance degradation when ablating *ei-*

Table 11. Full robustness results for modality noise injection, presented in two blocks. We report MAE(↓), Corr(↑), and F1(↑) as Gaussian noise (σ) is applied to Audio (A), Video (V), or both (A+V). Bold indicates the optimal performance (no noise).

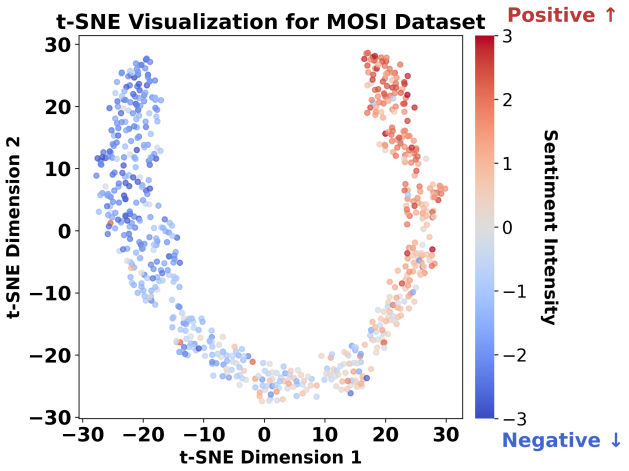
Noise		CH-SIMS			CH-SIMSv2		
		MAE ↓	Corr ↑	F1 ↑	MAE ↓	Corr ↑	F1 ↑
Full CICA	**0.00**	**0.378**	**0.754**	**85.68**	**0.245**	**0.842**	**85.89**
A	0.20	0.396	0.729	81.56	0.246	0.832	86.19
	0.40	0.407	0.752	85.70	0.284	0.798	86.09
	0.60	0.415	0.756	85.91	0.368	0.759	86.38
	0.80	0.423	0.775	85.68	0.462	0.582	71.69
V	0.20	0.438	0.636	75.61	0.257	0.819	86.28
	0.40	0.504	0.480	63.28	0.338	0.731	84.14
	0.60	0.532	0.397	47.94	0.460	0.487	65.96
	0.80	0.608	0.110	42.96	0.536	-0.406	30.97
A+V	0.20	0.446	0.635	75.59	0.260	0.818	85.99
	0.40	0.518	0.460	58.05	0.341	0.730	84.34
	0.60	0.548	0.369	53.17	0.460	0.487	66.15
	0.80	0.615	0.054	42.75	0.536	-0.410	30.52

Noise		MOSI			MOSEI		
		MAE ↓	Corr ↑	F1 ↑	MAE ↓	Corr ↑	F1 ↑
Full CICA	**0.00**	**0.630**	**0.855**	**88.14**	**0.489**	**0.856**	**85.15**
A	0.20	0.776	0.803	88.14	0.501	0.843	85.06
	0.40	0.970	0.702	77.54	0.516	0.808	84.87
	0.60	1.180	0.514	70.00	0.555	0.752	83.45
	0.80	1.397	0.100	44.67	0.618	0.662	76.20
V	0.20	0.791	0.799	88.14	0.497	0.822	84.32
	0.40	1.007	0.668	78.28	0.597	0.689	76.84
	0.60	1.223	0.453	63.35	0.754	0.405	57.65
	0.80	1.416	0.013	44.27	0.905	-0.221	34.51
A+V	0.20	0.802	0.795	87.84	0.505	0.817	84.72
	0.40	1.027	0.665	78.13	0.605	0.681	77.06
	0.60	1.236	0.439	63.35	0.757	0.400	56.97
	0.80	1.446	-0.113	38.19	0.911	-0.223	34.34

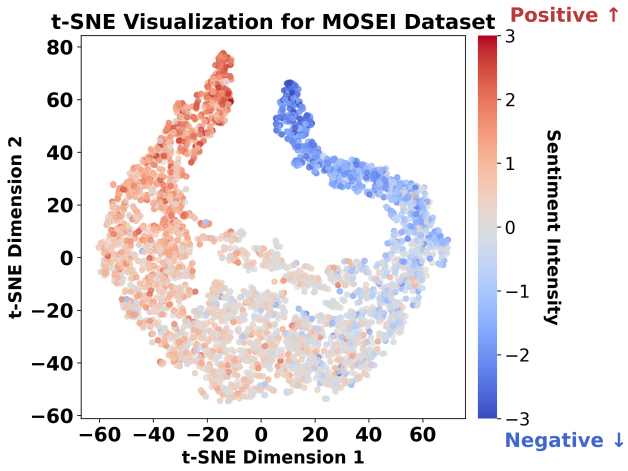
Table 12. Full robustness results for missing modality combinations across all four datasets. T+V+A denotes the full CICA model. Bold indicates the best performance for that metric on that dataset.

Modality Missing	CH-SIMS			CH-SIMSV2			MOSI			MOSEI		
	MAE	Corr	F1	MAE	Corr	F1	MAE	Corr	F1	MAE	Corr	F1
T+V+A	**0.378**	**0.754**	**85.68**	**0.245**	**0.842**	**85.89**	**0.630**	**0.855**	**88.14**	**0.489**	**0.856**	**85.15**
T+V	0.384	0.703	83.55	0.262	0.825	84.64	0.663	0.839	85.53	0.540	0.838	84.61
T+A	0.337	0.810	85.68	0.284	0.793	83.50	0.695	0.840	88.06	0.483	0.850	85.71
V+A	0.589	0.014	81.91	0.404	0.507	73.43	1.442	0.247	35.99	0.811	0.250	64.51
T	0.317	0.807	88.58	0.276	0.782	83.72	0.615	0.857	86.58	0.510	0.845	85.25
V	0.721	-0.089	46.90	0.441	0.477	67.99	1.435	0.094	33.18	0.811	0.232	66.00
A	0.591	0.032	81.91	0.549	0.123	73.21	1.433	0.254	40.17	0.839	0.124	57.39

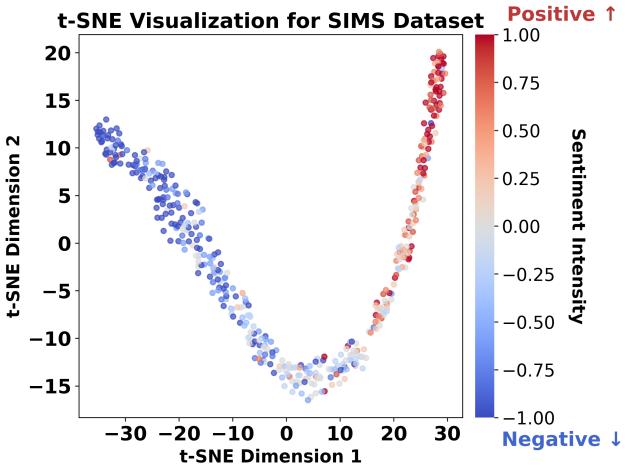
ther component provides rigorous, data-driven confirmation that s_m and u_m capture non-redundant, complementary aspects of reliability. Both signals are therefore required for the CICA framework to function.



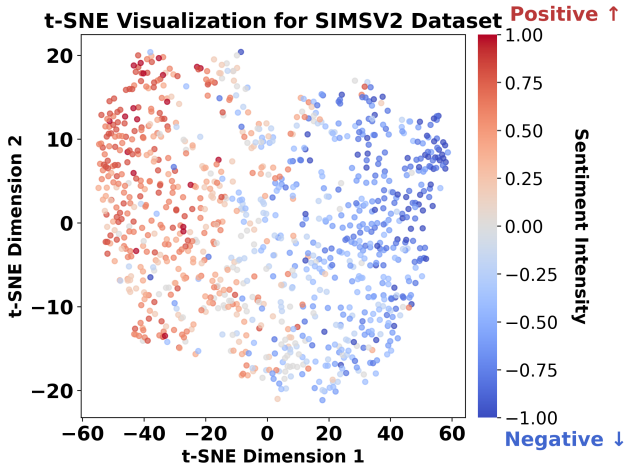
(a) MOSI



(b) MOSEI



(c) CH-SIMS



(d) CH-SIMSv2

Figure 6. **Full t-SNE visualization of CICA's fusion-layer embeddings across all four test sets.** Samples are colored by their ground truth sentiment polarity (*e.g.* Red: Positive, Blue: Negative, Green/Gray: Neutral). (a) **MOSI:** Shows clear separation, with a Pos-Neg centroid distance of 35.76. (b) **MOSEI:** A smooth manifold, validating the analysis in the main paper (Pos-Neg distance: 58.46). (c) **CH-SIMS:** Extremely tight and well-separated polar clusters (Pos-Neg distance: 40.74). (d) **CH-SIMSv2:** Demonstrates the strongest separation, with an inter-cluster distance of 66.20 between positive and negative centroids.

H.3. Principled Design and Verification of the Coupling Function

The coupling mechanism $r_m = g(s_m, u_m)$ links the perceived s_m and u_m signals to the final fusion weights. Our chosen formulation, $g(s_m, u_m) = \text{ReLU}(1 + s_m - u_m)$, was selected over alternatives for its theoretical and empirical advantages. It explicitly models the "net reliability margin" $(s_m - u_m)$ as a modulator applied to a baseline neutral weight of 1.0. This 'modulator' approach aligns with our "perceive-and-decide" paradigm: it amplifies reliable signals ($s_m > u_m$, $r_m > 1.0$), suppresses unreliable ones ($u_m > s_m$, $r_m < 1.0$), and preserves signals when $s_m \approx u_m$.

The ReLU activation is also important in this design. It no longer serves as a "sparse gate" at $s_m = u_m$, but rather as a "safety floor" to ensure the final weight remains non-negative. Complete suppression (truncation to zero) only occurs in cases of extreme unreliability (i.e., $u_m \geq s_m + 1$), preventing noise leakage. We empirically validated this design by comparing it against two plausible alternatives on the MOSEI dataset:

The results in Tab. 13 show that while all coupling mechanisms outperform the baseline (Variant D), our design achieves the best performance across all metrics. The superiority of our formulation over "Multiplicative" (Corr 0.856 vs. 0.853) stems from this explicit neutral baseline. The

$s_m \times (1 - u_m)$ gate lacks a clear "weight=1.0" point and can produce ambiguous scaling, especially when s_m or u_m are near 0.5. In contrast, our $1 + (s_m - u_m)$ design provides a strong inductive bias that explicitly modulates around the neutral 1.0 point. The MLP variant's slightly worse performance (Corr 0.850) likely stems from lacking this "net reliability margin" bias, which makes it harder to generalize the gating logic from limited data. These results confirm the $(1 + s_m - u_m)$ structure is a deliberate and essential component for our framework's robustness.

Table 13. Comparison of different coupling mechanisms on the MOSEI dataset. **No Coupling** corresponds to Variant D in Tab. 3. Bold indicates the best performance.

Function	Formula	MAE↓	Corr↑	Acc-2↑	F1↑
CICA (Ours)	$\text{ReLU}(1 + s_m - u_m)$	**0.489**	**0.856**	**84.72**	**85.15**
Multiplicative	$s_m \times (1 - u_m)$	0.492	0.853	84.55	84.98
Concat + MLP	$\sigma(\text{Linear}([s_m, u_m]))$	0.495	0.850	84.30	84.75
No Coupling	$r_m = 0$	0.504	0.820	83.51	83.82

I. Failure Cases and Sensitivity Analysis

To rigorously evaluate the potential risks of misguided confidence and understand the limitations of our framework, we conducted two additional analyses based on the MOSEI test set.

First, to test how strictly the Confidence-Informed Fusion (CIF) module adheres to the self-assessed reliability signals, we performed an Adversarial Confidence Perturbation test. By intentionally inverting the reliability scores generated by the CAP module during inference, we observed a severe performance degradation: the Pearson Correlation dropped significantly from 0.856 to 0.612, and the Mean Absolute Error (MAE) increased by 24%. This confirms that our fusion mechanism strictly follows the reliability guidance; when the confidence signals are corrupted, the fusion logic appropriately breaks down.

Second, we analyzed "Confidently Wrong" predictions—instances where the model assigned high confidence ($s_m > 0.8$) to its dominant modalities but still yielded a high prediction error. These failure cases are exceptionally rare, constituting only 3.4% of the test set. A qualitative review reveals that they are primarily driven by Deep Sarcasm. In such scenarios, the sensory signals (e.g., a bright smile or enthusiastic tone) are clear but intentionally deceptive. Because the signals are uncorrupted, the perception module confidently trusts them, causing the fusion module to be misled by the sarcastic facade. Addressing deep pragmatics and context-aware sarcasm reasoning remains an important avenue for future work.